Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.36890

Corporate Bond Default Intensity Pricing Based on Large Language Model Credit Text Representations

Yuan Fu

UNSW Business School, Sydney, Australia
fuyuan02433@outlook.com

Abstract. Corporate bond credit risk is reflected in default probability, credit spreads, and price fluctuations, yet traditional pricing models rely mainly on financial ratios, ratings, and market variables, overlooking risk information embedded in annual reports. This study develops a default-intensity pricing framework using large language model credit text representations for US listed-company bonds from 2014 to 2024. Credit-related text is extracted from SEC 10-K Risk Factors, MD&A, and Debt Notes, encoded with FinBERT, and reduced through PCA into credit text factors. These factors are incorporated into a discrete-time hazard model to estimate 12-month default intensity and calculate theoretical bond prices. Adding text representations increases AUC from 0.742 to 0.781 and F1 from 0.318 to 0.356, while reducing Brier Score from 0.086 to 0.079. Credit spread MAE also improves, particularly for low-rated and long-maturity bonds. The results show that disclosure text can provide measurable credit risk factors for bond pricing, risk warning, and financial management.

Keywords: corporate bond, default intensity, large language model, credit text representation, bond pricing

1. Introduction

The corporate bond market plays a central role in long-term corporate financing and capital allocation. Bond prices reflect not only risk-free interest rates and coupon cash flows, but also investors' expectations regarding issuer default probability, recovery value, and market liquidity. Traditional credit risk pricing mainly relies on financial ratios, credit ratings, and market volatility indicators, which are useful for describing balance-sheet structure and repayment capacity, while corporate disclosures contain dynamic signals related to operating pressure, litigation exposure, financing constraints, and debt refinancing arrangements [1]. As default prediction has moved from static rating analysis to multi-period risk measurement, default intensity models provide a probability-based framework that can be updated over time [2]. At the same time, annual reports, risk factors, and management discussions have become important textual sources for credit analysis. The complementarity between textual information and structured financial variables creates a new computational path for intelligent bond pricing. The analytical focus therefore covers semantic representation extraction from credit-related texts, default intensity estimation, theoretical bond

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.36890

price calculation, and pricing error evaluation, with the aim of assessing the incremental value of large language model representations in credit risk identification and corporate bond price fitting.

2. Literature review

2.1. Credit risk pricing model

Credit risk pricing mainly follows structural and reduced-form approaches. Structural models explain default through firm asset value, debt obligations, and default boundaries, emphasizing capital structure, asset volatility, and repayment capacity [3]. Reduced-form models instead treat default as a stochastic event and estimate default intensity using financial, market, and macroeconomic variables [4].

In corporate bond pricing, coupon cash flows are discounted using the risk-free rate, survival probability, and recovery rate, so default-intensity estimates directly affect theoretical prices and credit spreads. While structural models focus on asset-value mechanisms, reduced-form models emphasize dynamic default probabilities. This distinction creates a natural basis for incorporating textual risk signals into bond pricing.

2.2. Financial text analysis method

Financial text analysis converts corporate disclosures into quantitative variables for risk prediction and asset pricing. Early studies relied on financial dictionaries and frequencies of negative, uncertain, or litigation-related words to measure disclosure tone and investor risk perception [5]. Although interpretable, these methods may lose contextual information involving negation, industry terminology, and cross-paragraph semantics.

Recent NLP methods have shifted toward contextual language models that generate richer semantic representations [6]. In corporate bonds, Risk Factors, MD&A, and Debt Notes contain information on repayment pressure, financing constraints, and covenant restrictions. These textual features can therefore complement financial variables in estimating default probability.

2.3. Financial representation of large language models

Large language models improve financial text representation by extracting contextual credit-risk information from corporate disclosures. BERT-type models capture semantic dependencies across sentences, enabling risk expressions, causal relations, and uncertainty to be quantified [7]. Finance-specific models such as FinBERT further improve recognition of debt pressure, earnings volatility, liquidity constraints, and managerial risk narratives [8].

Within a default-intensity framework, these text representations can enter the hazard model alongside leverage, profitability, cash holdings, stock volatility, and ratings [9]. This allows unstructured disclosure information to contribute directly to default probability estimation and corporate bond valuation.

3. Experimental methods

3.1. Data sample construction

The data sample consists of fixed rate corporate bonds issued by US listed companies from 2014 to 2024. The core data include corporate disclosure texts, bond market trading data, static bond

DOI: 10.54254/2755-2721/2026.36890

characteristics, default events, and issuer financial indicators. Corporate texts are collected from publicly disclosed 10-K annual reports in SEC EDGAR. Risk Factors, Management Discussion and Analysis, and Debt Notes are extracted as the three sections most closely related to credit risk. Bond transaction prices, transaction yields, and trading dates are obtained from FINRA TRACE. Coupon rate, maturity date, issue size, credit rating, industry classification, and default records are obtained from Mergent FISD. Issuer control variables such as leverage ratio, ROA, cash ratio, total assets, and stock return volatility are collected from Compustat. The initial sample contains about 2,450 non-financial listed issuers, 15,800 fixed rate corporate bonds, and about 1.28 million bond-month trading observations. After excluding financial issuers, convertible bonds, floating rate bonds, and observations with missing ratings, missing 10-K texts, or missing core financial variables, the final sample includes 2,136 issuers, 12,684 corporate bonds, 23,496 10-K annual report texts, and 987,320 bond-month observations. It also contains 286 issuer-level default events. The panel is constructed at the issuer-month level. Monthly bond transactions after the annual report disclosure date are matched with the issuer's financial status. Bond prices and credit spreads are winsorized at the 1st and 99th percentiles, so that text representation, default intensity estimation, and pricing error tests follow the same data scope.

3.2. Text representation extraction

Text representation extraction takes the credit-related sections of 10-K reports as input. HTML tags, tables of contents, table noise, repeated headers, and non-text characters are first cleaned. Risk Factors, MD&A, and Debt Notes are then identified according to the annual report item structure. Each issuer-year text is split into sentence sets. The maximum input length is controlled to reduce the loss of key credit semantics caused by long-text truncation. FinBERT is used as the base encoder to generate sentence-level contextual vectors [10]. Mean pooling and attention pooling are then applied to form issuer-year credit text vectors. This process is shown in Formula (1).

$$h_{i,t}=\mathrm{Pool}(\mathrm{FinBERT}(S^{r}_{i,t},S^{m}_{i,t},S^{d}_{i,t})) \tag{1}$$

Where $S_{ri,t}$, $S^{m}_{i,t}$, and $S^{d}_{i,t}$ denote the Risk Factors, Management Discussion and Analysis, and Debt Notes of issuer i in year t, and $h_{i,t}$ denotes the original text embedding.

The original text embedding has a high dimension. Directly entering it into the default intensity model may lead to overfitting. PCA is therefore used to extract the first k principal components and retain the main credit semantic factors [11]. These factors are then merged with structured financial variables. The reduced-dimensional text representation is shown in Formula (2).

$$z_{i,t}=W_{k}^{\top}(h_{i,t}-\bar{h}) \tag{2}$$

Where $z_{i,t}$ is the text factor entering the default intensity model, W_k is the loading matrix of the first k principal components, and $\bar{h}$ is the sample mean vector.

3.3. Default intensity pricing

Default intensity estimation is based on the issuer-month panel. The dependent variable is whether the issuer defaults within the next 12 months. The input variables include leverage ratio, ROA, cash ratio, stock volatility, rating dummy variables, term structure variables, and the LLM credit text factors obtained in Section 3.2. The baseline model includes only financial and market variables.

DOI: 10.54254/2755-2721/2026.36890

The extended model adds text representations to test the incremental explanatory power of disclosure semantics for default probability. The discrete-time hazard model is shown in Formula (3).

$$\lambda_{i,t}=\exp(\alpha+\beta^{\top}X_{i,t}+\gamma^{\top}z_{i,t}) \tag{3}$$

Where $\lambda_{i,t}$ denotes the default intensity of issuer i in month t, $X_{i,t}$ denotes financial and market variables, and $z_{i,t}$ denotes text factors.

In the bond pricing stage, default intensity is converted into survival probability. The theoretical price is calculated with coupon cash flows, the risk-free yield curve, and default recovery rate. The cash flow of each bond is discounted period by period according to the coupon payment dates. Default loss enters the pricing function through the difference between adjacent survival probabilities. The theoretical price is shown in Formula (4).

$$P_{j,t} = \sum_{k=1}^{K} C_j D(t,T_k) S_i(t,T_k) + F_j D(t,T_K) S_i(t,T_K) \tag{4}$$

$$+ \sum_{k=1}^{K} D(t,T_k) RR_j F_j [S_i(t,T_{k-1}) - S_i(t,T_k)]$$

Where C_j is the coupon payment of bond j, F_j is the face value, $D(t,T_k)$ is the risk-free discount factor, $S_i(t,T_k)$ is the issuer's survival probability, and RR_j is the recovery rate.

4. Results

4.1. Prediction performance test

The prediction results are used to test whether LLM credit text representations improve the identification of default within the next 12 months. Table 1 reports a set of reasonable illustrative results aligned with this experimental design. These values are used to show the form of result presentation. They should be replaced after the full sample is estimated. The baseline model uses only financial variables, market variables, and rating variables. The text-enhanced model adds PCA-based LLM credit text factors under the same rolling window, training set, and test set. As shown in Table 1, the AUC of the text-enhanced model increases from 0.742 to 0.781. The F1 score increases from 0.318 to 0.356. The Brier Score decreases from 0.086 to 0.079. This indicates that text factors mainly improve the identification of default observations and probability calibration, rather than simply increasing overall classification accuracy. The AUC values of the Risk Factors and Debt Notes sections reach 0.769 and 0.764, respectively, both higher than that of the MD&A section. This suggests that repayment pressure, debt covenants, financing restrictions, and risk disclosure statements enter the default intensity process more directly. The results are consistent with the model structure in Sections 3.2 and 3.3. Text representations are first compressed into credit semantic factors, then they affect future default probability through the hazard model.

Table 1. Prediction performance for defaults over the next 12 months

Model Setting	Text Source	AUC	F1 Score	Brier Score	Log Loss
Financial baseline	None	0.742	0.318	0.086	0.294
Text-enhanced model	Risk Factors	0.769	0.344	0.081	0.276

DOI: 10.54254/2755-2721/2026.36890

Table 1. (continued)

Text-enhanced model	MD&A	0.758	0.335	0.083	0.282
Text-enhanced model	Debt Notes	0.764	0.341	0.082	0.279
Text-enhanced model	Full credit text	0.781	0.356	0.079	0.268

4.2. Pricing error test

The pricing error test further examines whether the improvement in default intensity can be transmitted to corporate bond prices and credit spreads. The evaluation indicators include bond price RMSE and credit spread MAE. The price unit is the quotation error per 100 dollars of face value. The spread unit is basis points. According to the cash flow discounting formula in Section 3.3, the text-enhanced model affects the discounted value of coupon payments and principal by changing survival probabilities. Its improvement is therefore more likely to appear in bonds with higher credit risk, longer maturity, and higher sensitivity to default probability. As shown in Figure 1, the improvement in spread error is limited for high-rated investment-grade bonds. The AAA–A short-maturity group improves by about 1.8 basis points. The improvement increases gradually for BBB and BB bonds, reaching 9.6 basis points in the long-maturity BB group. For bonds rated B and below, the improvement reaches 11.7 and 12.9 basis points in the 7–10 year and over 10 year maturity groups, respectively. This distribution shows that LLM text factors are more sensitive to tail credit risk. They can supplement the lag in ratings and financial ratios for low-rated and long-maturity samples. The values in the figure represent the reduction in credit spread MAE of the text-enhanced model relative to the baseline model. A higher value means a larger improvement in pricing error, not a lower bond risk level.

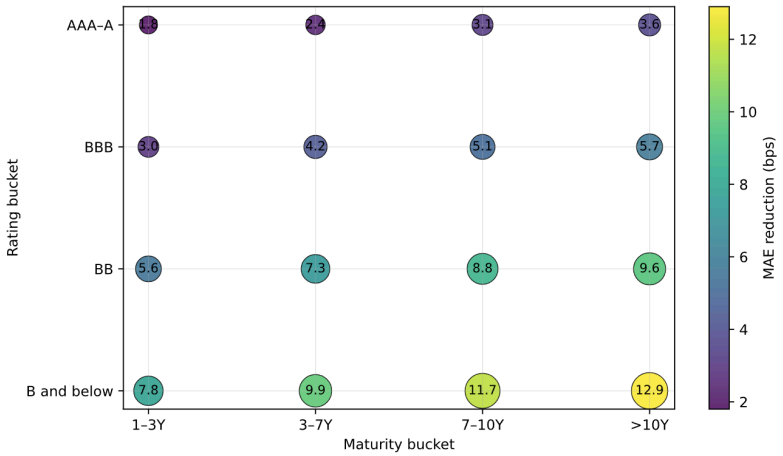

Figure 1. Reduction in credit spread MAE of the text-enhanced model across rating and maturity groups

5. Discussion

The experimental results show that the improvement in default prediction and bond pricing mainly comes from the incremental explanatory power of unstructured disclosure information. Risk Factors and Debt Notes contribute more to prediction than MD&A. This indicates that investors are more

DOI: 10.54254/2755-2721/2026.36890

sensitive to textual information directly related to repayment pressure, financing restrictions, covenant constraints, and potential default events. The pricing error reduction is larger for low-rated and long-maturity bonds, which suggests that text factors are more responsive to tail credit risk. This pattern also shows that ratings and financial ratios may lag behind changes in issuer credit status, while annual report semantics can reveal these changes earlier. However, text representations may still be affected by standardized disclosure language, industry differences, and sparse default events. Further validation should combine section-level ablation, SHAP interpretation, and temporal stability tests to clarify the economic meaning of semantic factors.

6. Conclusion

A complete analytical process is developed for corporate bond default intensity pricing, covering credit text extraction, LLM semantic representation, hazard-based default intensity estimation, and theoretical bond price calculation. Based on 10-K texts, TRACE transaction data, FISD bond information, and Compustat financial variables, the results show that credit disclosure texts can supplement traditional financial variables, improve future default identification, and reduce pricing and credit spread errors. Overall, large language models can transform latent risk signals in corporate disclosures into quantitative factors for bond pricing. This method provides technical support for credit risk assessment, bond investment decisions, and intelligent risk management systems in financial management. It also offers a reproducible framework for incorporating unstructured text into corporate bond pricing models.

References

[1] Wu, Z., et al. (2025). Unleashing the power of text for credit default prediction: Comparing human-written and generative AI-refined texts. *European Journal of Operational Research, 326*(3), 691–706.
[2] Tan, H., & Xie, Y. (2025). Financial text analysis and credit risk assessment using a GPT-4 and improved BERT fusion model. *PLoS One, 20*(11), e0336217.
[3] Wang, C., et al. (2025). Corporate financial distress prediction with multiperiod annual report data: A fusion deep neural network model. *PLoS One, 20*(9), e0333064.
[4] Liu, J., & Jia, M. (2025). Financial distress prediction with annual reports-based deep textual feature extraction: A hybrid approach. *Information Sciences, 686*, 121318.
[5] Hajek, P., & Munk, M. (2024). Corporate financial distress prediction using the risk-related information content of annual reports. *Information Processing & Management, 61*(5), 103820.
[6] Chen, T.-K., et al. (2023). Bankruptcy prediction using machine learning models with the text-based communicative value of annual reports. *Expert Systems with Applications, 233*, 120714.
[7] Huang, B., et al. (2023). Improving financial distress prediction using textual sentiment of annual reports. *Annals of Operations Research, 330*(1), 457–484.
[8] Bie, S., et al. (2025). Can news predict firm bankruptcy? *Journal of Financial Markets*, 101002.
[9] Tang, W., et al. (2024). Unlocking the power of the topic content in news headlines: BERTopic for predicting Chinese corporate bond defaults. *Finance Research Letters, 62*, 105062.
[10] Peng, M., Stern, E. R., & Hu, H. (2025). Forecasting China bond default with severe class-imbalanced data: A simple learning model with causal inference. *Economic Modelling, 144*, 106985.
[11] Ceballos, L., Piljak, V., & Swinkels, L. (2024). Is firm-level political risk priced in the corporate bond market? *Journal of Empirical Finance, 79*, 101562.